INTERPRETABILITY AND PERFORMANCE ANALYSIS OF DEEP NEURAL STRUCTURES ON TIME SERIES CLASSIFICATION TASK

Prepared by

Berk GORGULU 1005079898

Shiva SAXENA 1005609370

INTERPRETABILITY

- "understanding", "interpreting", or "explaining".
- Post-hoc interpretability,
  - i.e. a trained model is given and our goal is to understand what the model predicts (e.g. categories) in terms what is readily interpretable (e.g. the input variables).
- **Definition 1.** ***Interpretation*** is the mapping of an abstract concept (e.g. a predicted class) into a domain that the human can make sense of.
- **Definition 2.** ***Explanation*** is the collection of features of the interpretable domain, that have contributed for a given example to produce a decision (e.g. classification or regression).
  - A heatmap highlighting which pixels of the input image most strongly support the classification decision.

Areas : Finance, healthcare and self-driving cars

Problems : Critical Decisions, discriminatory (e.g. gender based) or violation of laws.

Based on Lapuschkin et al. (2016) "Analyzing classifiers: Fisher vectors and deep neural nets"

TIME SERIES CLASSIFICATION

Motivation :

- Important applications including finance, science, medicine and multimedia.

Main Challenges :

- Finding a suitable representation with reduced dimensionality
- Discovering the important patterns

Instance-Based Methods

- Sensitive to noise, scaling, translation and dilation of the patterns
- Computational and memory inefficiencies
- Lack of interpretability

Feature-Based Methods

- Feature engineering is hard in time-series domain
- Can't identify the pattern shifts

Shapelet-Based Methods

- Midway approach

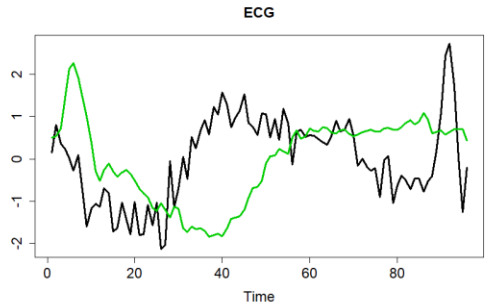


Figure 2. ECG of healthy and unhealthy individuals

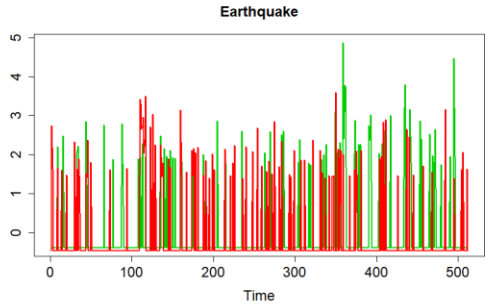
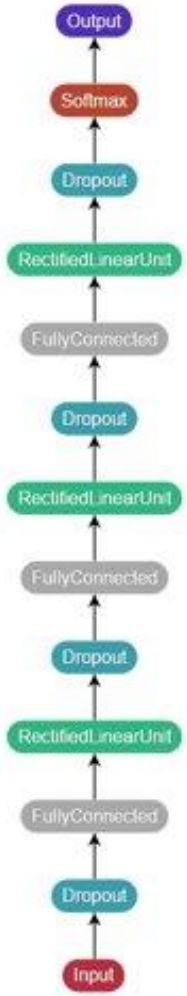


Figure 3. Seismic waves for predicting an earthquake will ocur or not

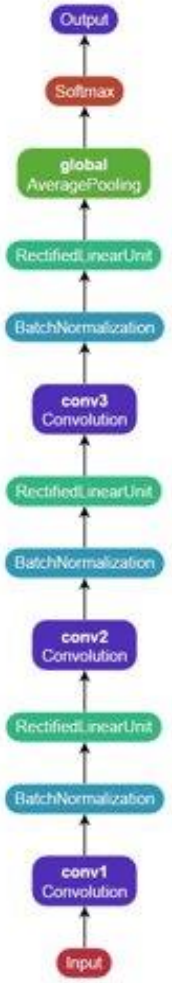
CANDIDATE MODELS

(WANG ET AL., TIME SERIES CLASSIFICATION FROM SCRATCH WITH DEEP NEURAL NETWORKS: STRONG BASELINE)

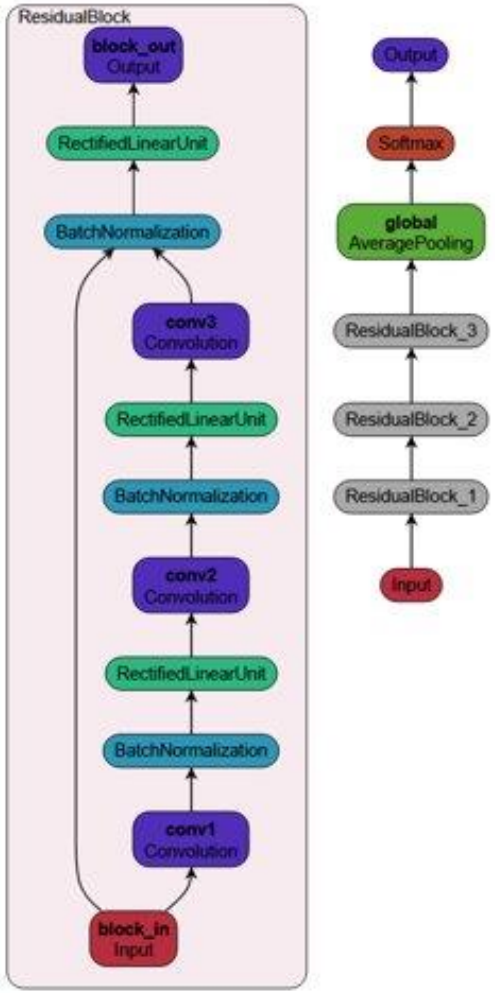
MULTI-LAYER PERCEPT



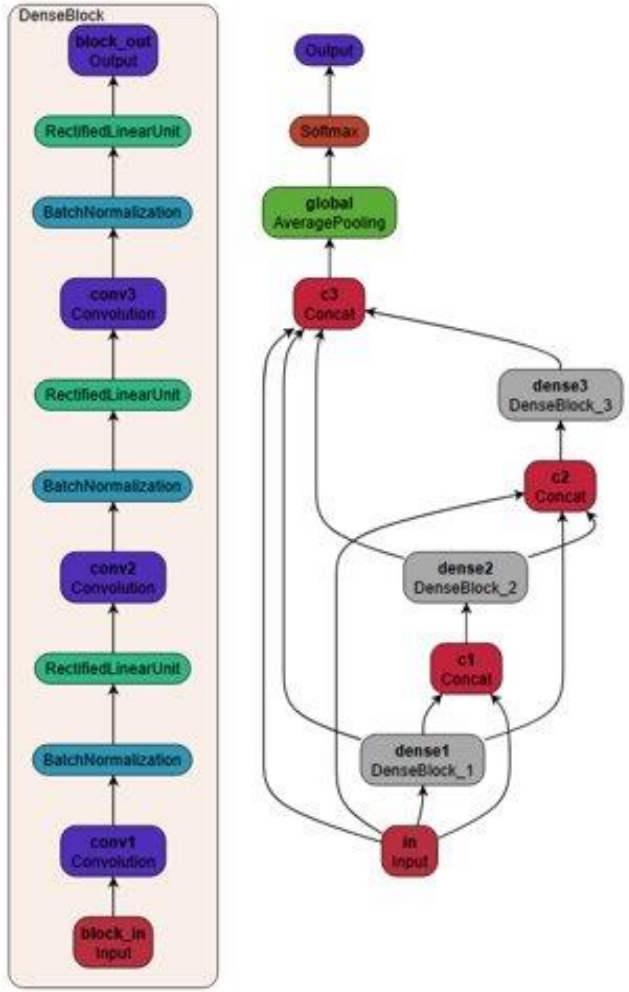
DEEP CONV NETS



RESIDUAL NETWORKS



DENSELY CONNECTED NETS



PERFORMANCE COMPARISONS

	MLP	Deep CNN	ResNet	DenseNet
Beef	**0.867**	0.8	0.833	0.767
CBF	0.86	**1**	**1**	0.99
Coffee	**1**	**1**	**1**	**1**
Gun Point	0.953	**1**	0.993	**1**
ShapesAll	0.753	0.835	**0.885**	0.877
Strawberry	0.96	0.965	**0.970**	0.964
Two Patterns	0.888	0.897	0.948	**0.956**
Wafer	0.995	**0.996**	0.993	0.995
Yoga	0.852	0.865	**0.874**	**0.874**

Methods for Interpreting and Understanding Deep Neural Networks

Grégoire Montavon[a,*], Wojciech Samek[b,*], Klaus-Robert Müller[a,c,d,*]

[a] *Department of Electrical Engineering & Computer Science, Technische Universität Berlin, Marchstr. 23, Berlin 10587, Germany*
[b] *Department of Video Coding & Analytics, Fraunhofer Heinrich Hertz Institute, Einsteinufer 37, Berlin 10587, Germany*
[c] *Department of Brain & Cognitive Engineering, Korea University, Anam-dong 5ga, Seongbuk-gu, Seoul 136-713, South Korea*
[d] *Max Planck Institute for Informatics, Stuhlsatzenhausweg, Saarbrücken 66123, Germany*

Abstract

This paper provides an entry point to the problem of interpreting a deep neural network model and explaining its predictions. It is based on a tutorial given at ICASSP 2017. It introduces some recently proposed techniques of interpretation, along with theory, tricks and recommendations, to make most efficient use of these techniques on real data. It also discusses a number of practical applications.

Keywords: deep neural networks, activation maximization, sensitivity analysis, Taylor decomposition, layer-wise relevance propagation

INTERPRETATION TECHNIQUES

- Sensitivity Analysis
- Taylor Approximation
- Relevance Propogation

SENSITIVITY ANALYSIS

- A first approach to identify the most important input features is sensitivity analysis.
- It is based on the model's locally evaluated gradient or some other local measure of variation.

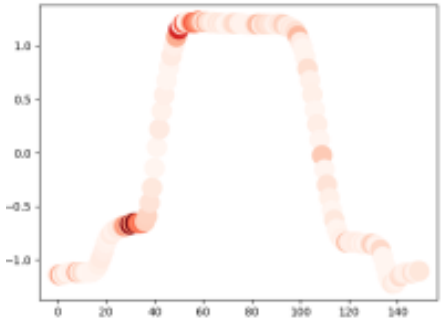
$$R_i(\boldsymbol{x}) = \left(\frac{\partial f}{\partial x_i}\right)^2$$

- The most relevant input features are those to which the output is most sensitive.
- Sensitivity scores are a decomposition of the local variation of the function as measured by the gradient square norm:

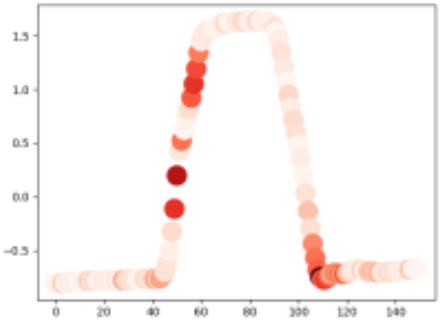
$$\sum_{i=1}^{d} R_i(\boldsymbol{x}) = \|\nabla f(\boldsymbol{x})\|^2$$

- Sensitivity analysis does not produce an explanation of the function value f(x), but rather a variation of it
  - e.g. detecting cars in images, we answer the question "what makes this image more/less a car?", rather than the more basic question "what makes this image a car?".

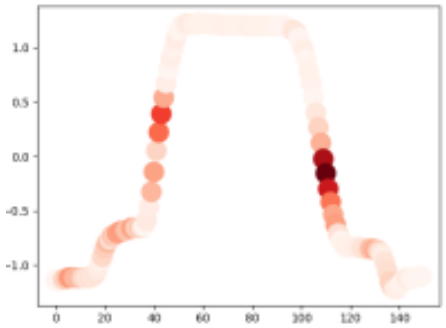
SENSITIVITY ANALYSIS



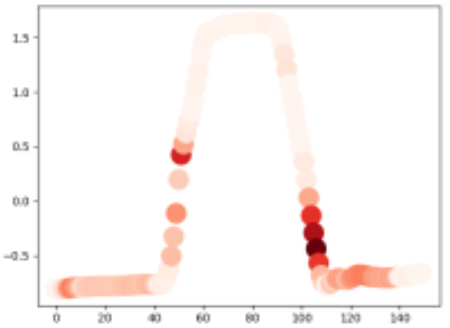
(a) MLP Class 1



(b) MLP Class 2



(c) CNN Class 1



(d) CNN Class 2

SIMPLE TAYLOR DECOMPOSITION

- Explains the model's decision by decomposing the function value $f(x)$ as a sum of relevance scores.

$$f(\boldsymbol{x}) = \sum_{i=1}^{d} R_i(\boldsymbol{x}) + O(\boldsymbol{x}\boldsymbol{x}^\top)$$

where the relevance scores

$$R_i(\boldsymbol{x}) = \frac{\partial f}{\partial x_i}\Big|_{\boldsymbol{x}=\widetilde{\boldsymbol{x}}} \cdot (x_i - \widetilde{x}_i)$$

- Because these higher-order terms are typically non-zero, this analysis only provides a partial explanation of $f(x)$.

- A special class of functions, piecewise linear and satisfying the property $f(tx) = t\,f(x)$ is not subject to this limitation.
  - Homogeneous linear models, or deep ReLU networks (without biases)

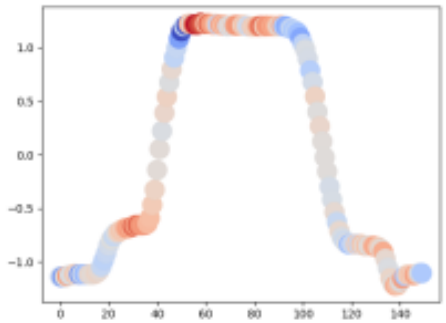
$$f(\boldsymbol{x}) = \sum_{i=1}^{d} R_i(\boldsymbol{x})$$

where the relevance scores simplify to

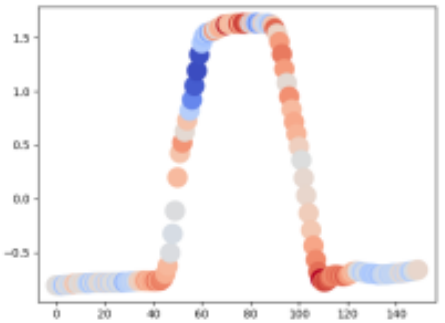
$$R_i(\boldsymbol{x}) = \frac{\partial f}{\partial x_i} \cdot x_i.$$

- Relevance = sensitivity (partial derivative) * saliency (input value).
- That is, an input feature is relevant if it is both present in the data, and if the model reacts to it.

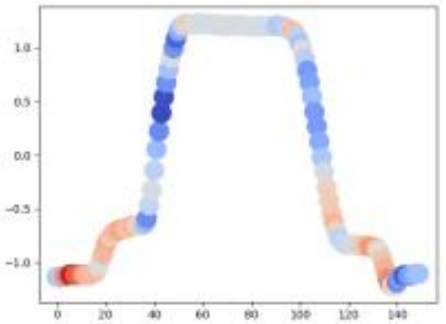
SIMPLE TAYLOR APPROXIMATION



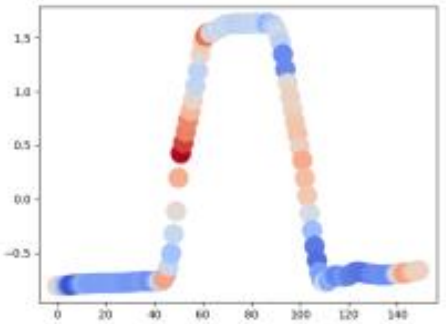
(a) MLP Class 1



(b) MLP Class 2



(c) CNN Class 1

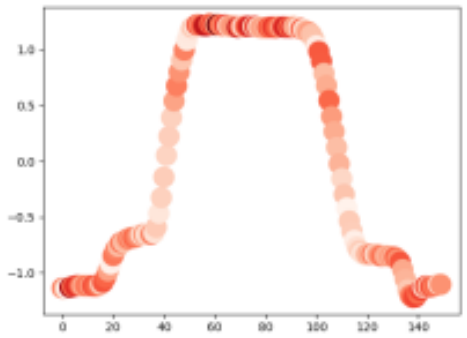


(d) CNN Class 2

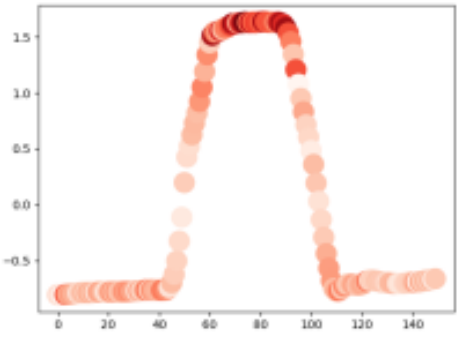
LRP (LAYER-WISE RELEVANCE PROPAGATION)

- LRP provides an interpretation methods that uses the feed-forward graph structure.
- The algorithm starts at the output of the network, and moves in the graph in reverse direction, progressively redistributing the prediction score (or total relevance) until the input is reached.
- Relevance score is conserved in each layer and distributed to previous layer according to its activation and weights.
- The network can identify the important parts of the inputs.

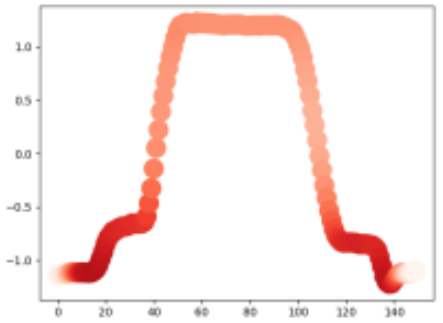
LRP (LAYER-WISE RELEVANCE PROPAGATION)



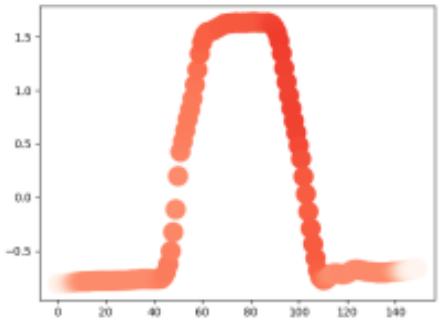
(a) MLP Class 1



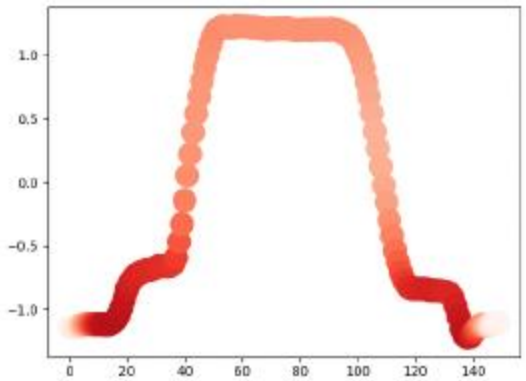
(b) MLP Class 2



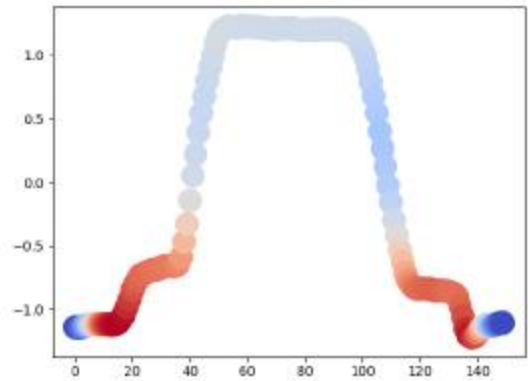
(c) CNN Class 1



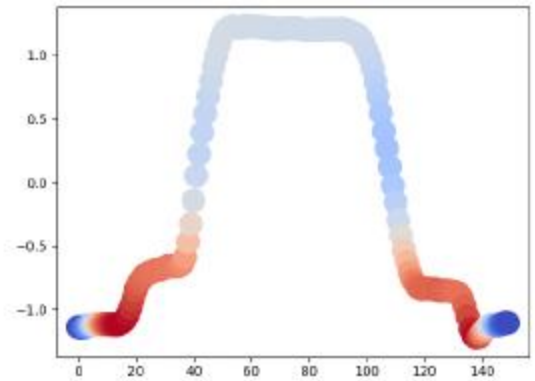
(d) CNN Class 2



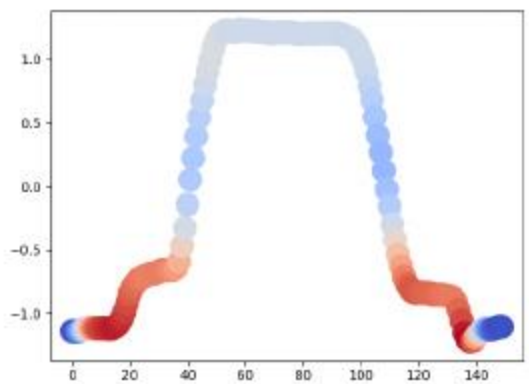
(a) Class 1, $\alpha = 1$



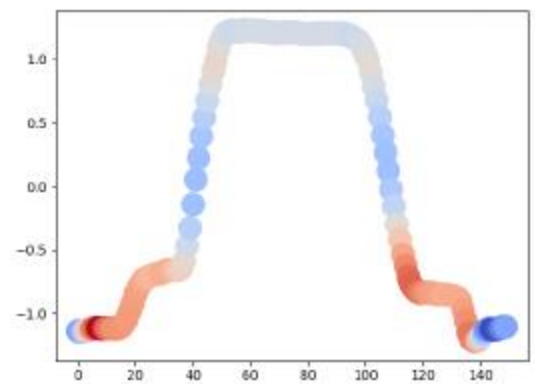
(b) Class 1, $\alpha = 2$



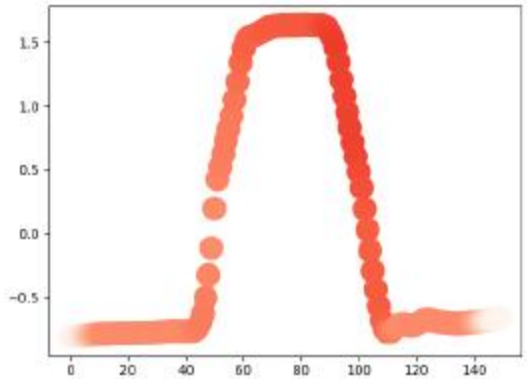
(c) Class 1, $\alpha = 5$



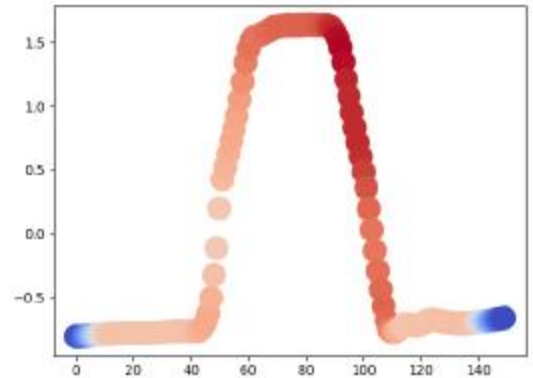
(d) Class 1, $\alpha = 10$



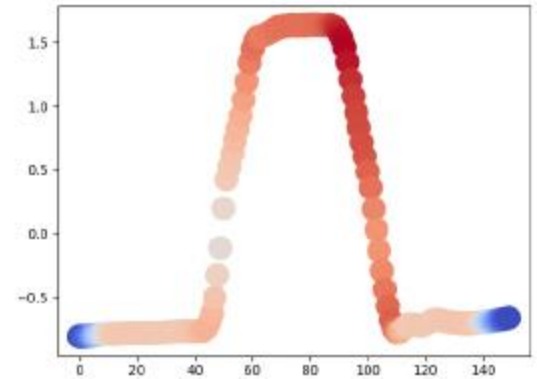
(e) Class 1, $\alpha = 20$



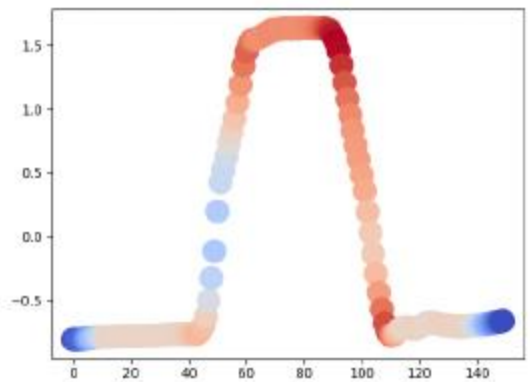
(f) Class 2, $\alpha = 1$



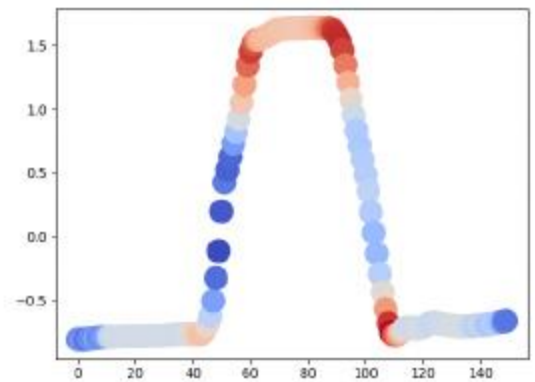
(g) Class 2, $\alpha = 2$



(h) Class 2, $\alpha = 5$



(i) Class 2, $\alpha = 10$



(j) Class 2, $\alpha = 20$

CONCLUSION

- Deep Convolutional Networks are more suitable network structures for time series classification task in terms of performance.
- Performance of the Deep Convolutional Networks can be improved by using Residual Networks or Densely Connected Networks
  - Due to the 1-D nature of the data, these kind of networks do not make the differences they provide in the image classification
- Deep Convolutional Networks provide the most interpretable model
- LRP provides the best interpretation tool among the other alternatives tried in this work.

  •Although it is mainly used for image classification task, it performs very well in 1-D time series classification.